Statistisches Parsing

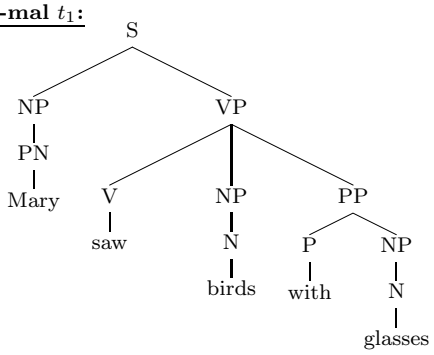
Detlef Prescher

June 19, 2007

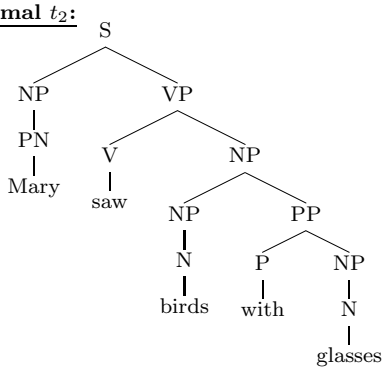
Hausaufgabe 8

(i) Benutze Charniaks unlexikalisiertes Modell, um mit der folgenden Baumbank eine unlexikalisierte PCFG zu trainieren:

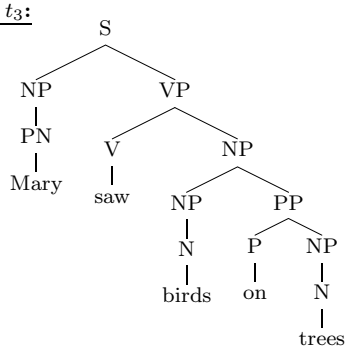
100-mal t_1:



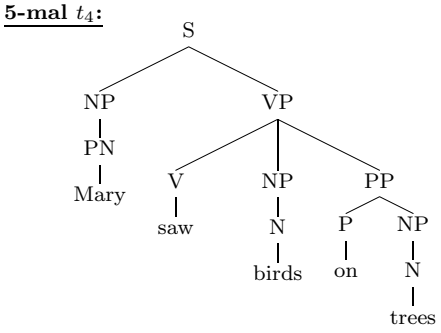
5-mal t_2:



100-mal t_3:



5-mal t_4:



Zeige, dass zwar $p(t_1) > p(t_2)$, jedoch leider auch $p(t_3) < p(t_4)$ gilt.

(ii) Beweise, dass jede probabilistische Version der zugrundeliegenden symbolischen Grammatik sich derart schlecht verhält.

(iii) Lexikalisiere die Baumbank in (i) mit Hilfe der folgenden Kopfmarkierungen:

```
S  -> NP VP'
```

```
VP -> V' NP PP
VP -> V' NP
NP -> NP' PP
PP -> P' NP
```

(Die unären Regeln benötigen keine Kopfmarkierungen.)

(iv) Benutze Charniaks lexikalisiertes Modell, um mit der Baumbank in (iii) eine lexikalisierte PCFG zu trainieren (ohne Smoothing für die lexikalisierten Regeln und Dependencies). Welche Disambiguierungsergebnisse erhälst Du nun?

(v) Benutze Collins lexikalisiertes Modell, um mit der Baumbank in (iii) eine lexikalisierte PCFG zu trainieren (nur Markovisierung von Regeln mittels STOP-Symbolen). Welche Disambiguierungsergebnisse erhälst Du nun?

Bemerkung: Dieses Aufgabenblatt ist prüfungsrelevant.

Abgabetermin: Dienstag, 26. Juni

Besprechung: Donnerstag, 28. Juni